Healthcare Cost Predictor – Neural Network

Final Report

Problem Statement

Healthcare is a major expense across the world. Particularly, in the United States, the CDC estimated national healthcare expenditures to have reached 3.8 trillion in 2019. This enormous cost is largely passed onto the patient and their health insurance provider. Estimating individual health care costs based on various demographic and health features is an important step in informing both individuals of their projected healthcare costs, and government programs on who best to target for financial assistance.

Therefore, I sought to design a deep learning neural network to predict individual healthcare expenses. The goal was to determine which factors contributed most heavily to healthcare cost and to minimize the prediction error of the model.

Data Wrangling

To build my prediction model, I selected a healthcare dataset on Kaggle that can be found at this link. The Dataset included 1,338 rows (or user-cost instances) and the seven columns shown below:

Age	The insured person's age
Sex	Gender (male or female) of the insured
BMI	(Body Mass Index): A measure of body fat based on height and weight
Children	The number of dependents covered
Smoker	Whether the insured is a smoker (yes or no)
Region	The geographic area of coverage
Charges	The medical insurance costs incurred by the insured person

It is important to note that there is some ambiguity on the unit of currency the charges column shows, what period of time these costs were incurred, and the specific location they were incurred. However, these details are not particularly important for the building of this model, as it is simply intended to be used as a demo model for future development. The first step in cleaning my data was to check for null values and duplications. The dataset was very clean with zero null values and only one duplicate which I resolved by simply dropping one of the copies.

Before proceeding to the next stage of exploratory data analysis, I wanted to create a heatmap to determine which features correlated most with charges. I decided to go ahead and conduct preprocessing on my dataset in order to generate this. The first step was to do

one hot encoding on my categorical features. This included the 'sex' and 'smoking' columns which were binary, and the 'region' column which gave rise to three new columns, each representing a separate region. I then decided to apply the StandardScaler object from sklearn on all my non-binary features. I now had two dataset: one for EDA and another one for generating my heatmap and using on my upcoming neural network model.

Exploratory Data Analysis

As previously stated, I wanted to get a sense of the correlative relationship between my features and the target variable before I visualized my data using more classical techniques. As fig. 1 shows, the top three strongest correlators with charges were smoking, age, and body mass index.

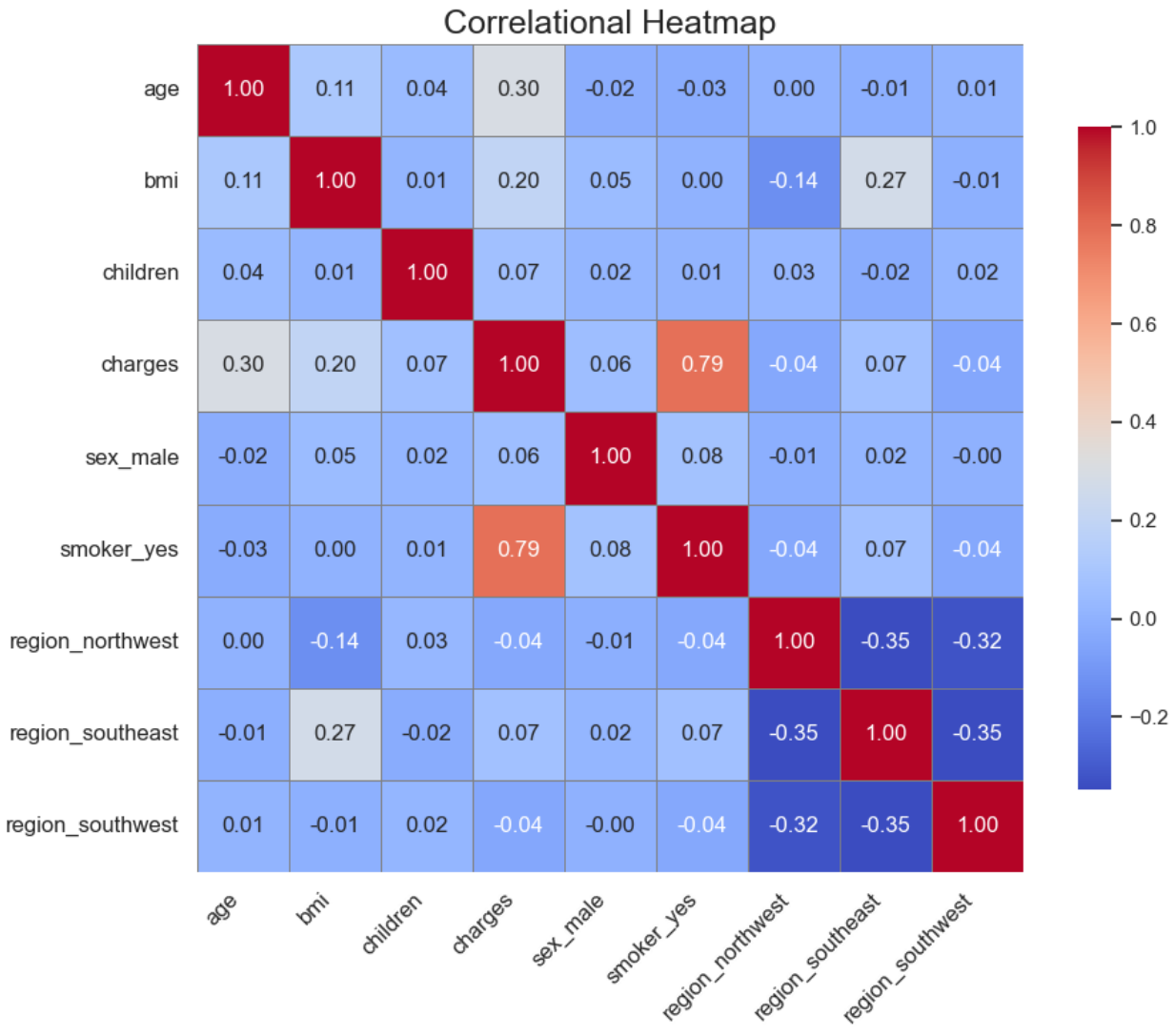

Figure 1: Heatmap showing Pearson correlation coefficient values between all features in model dataset

These strong positive correlations were further confirmed by examining a few different scatter plots. In particular, the relative correlation of the smoking feature on charges compared to age and BMI was very great as fig. 2 displays. While charges were higher but

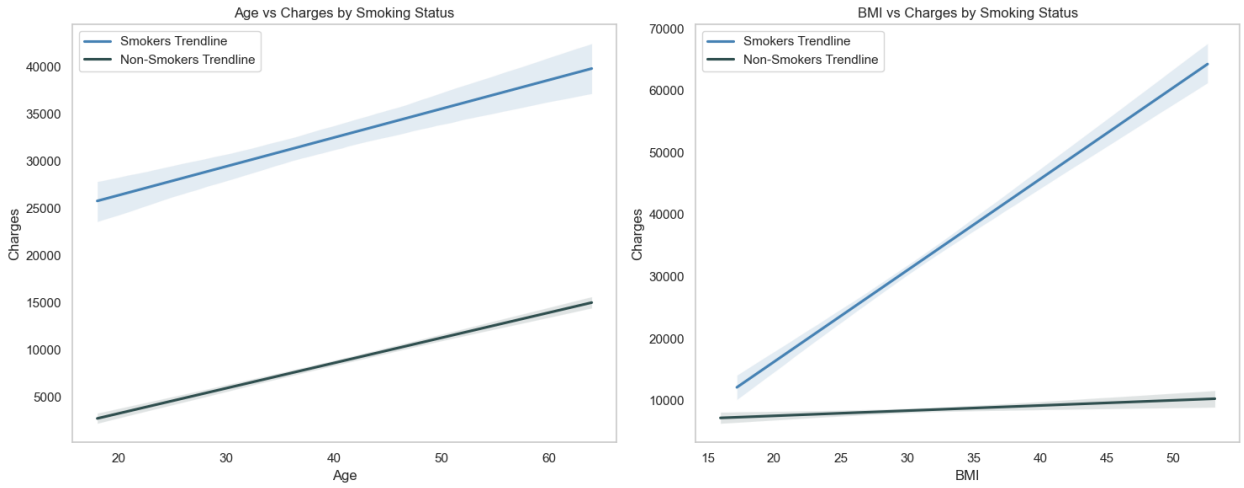

Figure 2: Scatterplot trendlines illustrating the differential correlation between smoking status and charges based on age and BMI

increased at an almost identical rate for smokers with respect to increasing age, individuals who smoked and had higher BMI experienced increasingly higher charges.

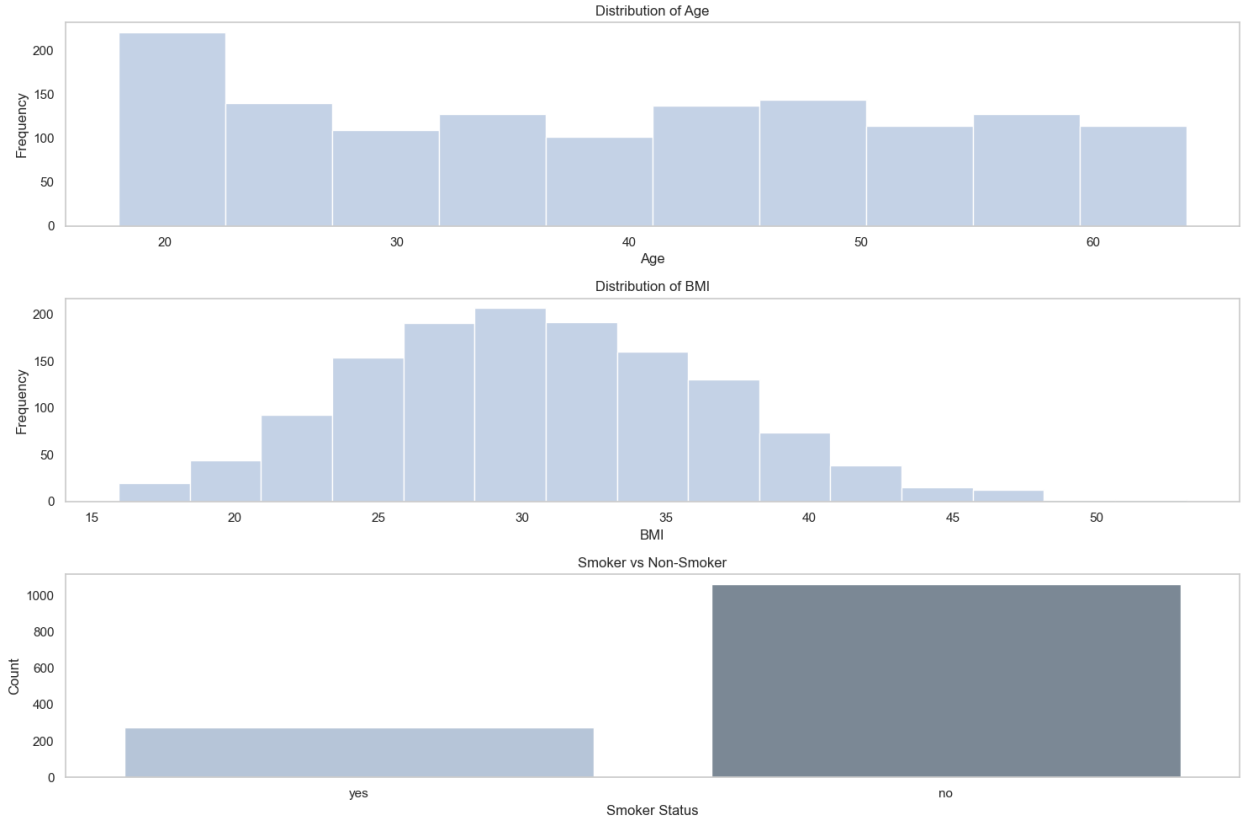

Figure 3: Distribution of age, BMI, and smoking status

Before beginning the modeling phase, I aimed to understand the distribution of my top three correlates, which are illustrated in Figure 3. The age distribution in the dataset was relatively balanced, with a slight increase in the number of younger individuals. The Body Mass Index (BMI) exhibited a normal distribution centered around 30. Additionally, the smoker status revealed that there were more than twice as many non-smokers in the dataset compared to smokers. With these distributions aligning well with what one might expect, I felt confident and prepared to move forward into the modeling phase.

Modeling

Considering I already had one hot encoded and standardized my model dataset for the correlational heatmap, I was ready to begin building my neural network. However, I wanted to generate a baseline regression model using sklearn's Random Forest Regressor to compare against future neural networks. I conducted an 80/20 split and used 100 estimators for my baseline RF model. After generating predictions, I inverse scaled the tested and predicted charge values to better interpret the meaning of my performance metrics. I did this for all subsequent models. This gave me a mean absolute error (MAE) of 2,934.18 and an R-squared value of 0.80. This indicated that on average the RF model predicted the charges +/- $2934.18 (assuming the units were USD) of the actual value. Moreover, the R-squared indicates that 80% of the variance of my dataset is accounted for with this model.

I was now prepared to build a series of three neural networks with different architectures to see how they would compare against my baseline model. Each model had four fully connected dense layers using the ReLU activation function, except the last one, which had no activation to allow continuous output for predicting the target variable. All networks used the Adam optimizer and mean squared error (MSE) was used as the loss function. To prevent overfitting, early stopping was set to halt training if validation performance didn't improve for three epochs, meaning the model could stop before the maximum of 200 epochs. The only parameter that differed between my models was the number of nodes at the first three layers.

For the first neural network (NN), I used a 64, 32, 16 node structure for my first three layers respectively. For the second NN, I used a 100, 50, 25, and for the third NN, I used a 50, 100, 50 node structure. The third NN had the best mean absolute error of 3,055.17 and

Comparison of Neural Network Models

Model	MAE	R^2
NN1 (64,32,16)	3329.68	0.8
NN2 (100,50,25)	3384.37	0.79
NN3 (50,100,50)	3055.17	0.79

Figure 4: Table Showing Three Neural Network Model Performance Outcomes

had a comparable R^2 to the baseline model of 0.79. However, even the best NN did not perform as well as my baseline RF model.

I decided to conduct a best parameter search on my NN using the hyperband tuner from Keras. This gave an optimal node structure of 72, 120, 40 for the first three layers of my NN. It also showed an optimal learning rate of 0.001. The remaining parameters were the same as my previously built NNs. I went ahead and built the model. It gave the following metrics:

Mean Absolute Error: 2808.8756

R^2 Score: 0.7882

This model performed better on MAE and only marginally worse in capturing the variability of the dataset compared to all previous models. However, the magnitude of

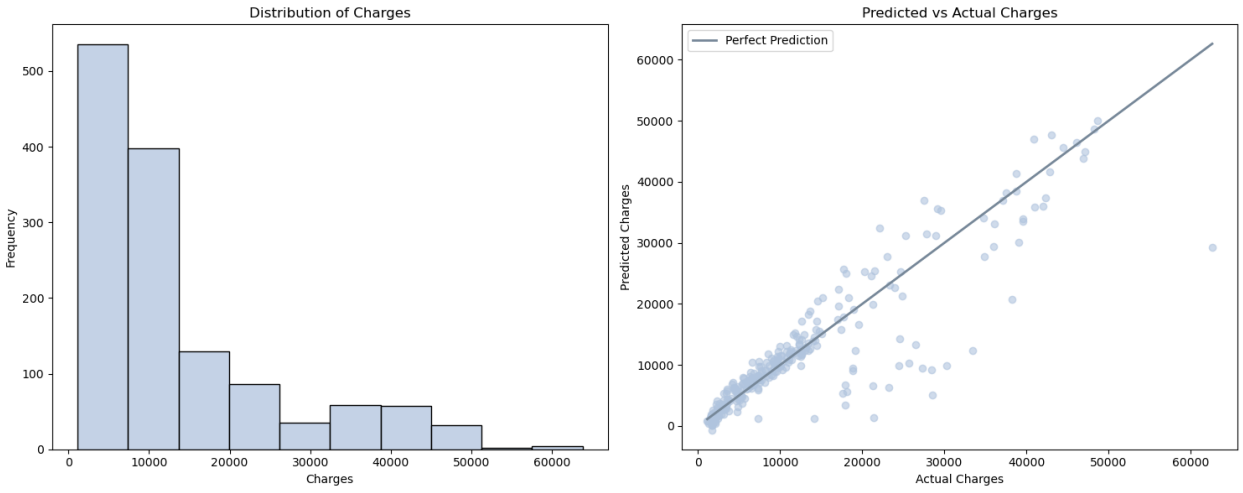

Figure 5: Histogram of distribution of charges & scatterplot of predicted vs actual charges

improvement was not that significant across any of the models. I suspect that this was due to haveing a relatively small dataset for the NN to train on with only a small sample of indivduals at the upper end of the charges distribution. This is illustrated well in figure 5 which shows predicted vs actual charges on the right. As you can see, the model fits well for charges $15,000 and below where there is a large sampling of charges. However, as the charges begin to increase and the sampling of charges decrease as the histogram in figure 5 on the left shows, the fit of the model worsens.

In addition to these consideration, I was interested in seeing if the feaures I identified as being the highest correlates to cost were also the most import features to the model. As figure 6 shows, smoking, age, and BMI were inded the top three most import features for model performance and are valuable features to include in future modeling.

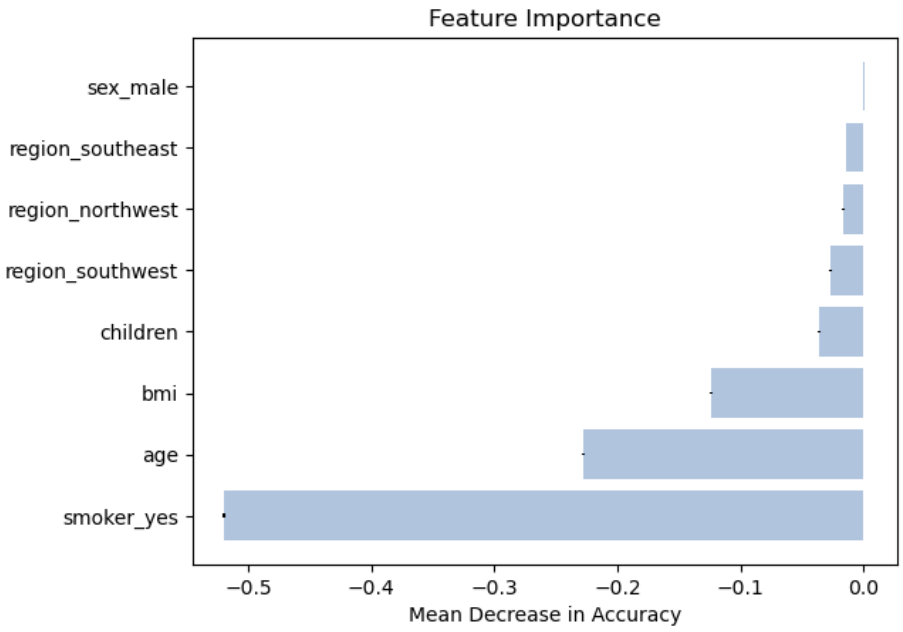

Figure 6: Bar plot showing relative importance of features for model prediction

Application & Future Work

While improving model performance significantly has proved challenging, it is essential to recognize the predictive capabilities of the models. The dataset shows that healthcare charges range from approximately $1,000 to $65,000. With a mean absolute error of around $3,000, the models can predict healthcare costs within a margin of ±$3,000. This level of accuracy can be quite valuable for individuals and organizations looking to estimate potential healthcare expenses.

To enhance future model development, a couple of steps can be taken:

1. Increase the sample size to ensure a sufficiently large pool of high charges for training.

2. Explore other modeling approaches beyond deep learning, such as XGBoost or Random Forest, with additional parameter tuning.

Implementing these strategies could help reduce prediction errors and further improve model performance.